The Geometry of Fragility: Benchmarking Robustness of Time Series Explanations via Curvature Maximization

Yueshan Chen*, Sihai Zhang†
†School of Integrated Circuits, University of Science and Technology of China, Hefei, Anhui, P.R. China
†School of Cyberspace Science and Technology, University of Science and Technology of China
Email: cys515@mail.ustc.edu.cn, shzhang@ustc.edu.cn

***Abstract*—Explainable Artificial Intelligence is critical for trustworthy time series forecasting, yet recent studies reveal its vulnerability to adversarial manipulations. Existing attacks, primarily designed for images, fail to preserve temporal semantics and often require white-box access, rendering them ineffective against non-differentiable methods like LIME and SHAP. To bridge this gap, we propose the Iso-surface Curvature Maximization Attack, a geometric attack method grounded in the insight that explanation instability is intrinsically governed by the curvature of the prediction manifold. Instead of targeting specific explainer outputs, ICMA directly maximizes the local curvature of the decision boundary under a spectro-temporal constraint. This constraint strictly confines perturbations to the high-frequency residual subspace, preserving the underlying trend and seasonality. Extensive experiments on real-world datasets across diverse architectures confirm that maximizing this geometric objective significantly destabilizes explanations, empirically verifying that high curvature is indeed the primary driver of explanation fragility.**

***Index Terms*—Time Series, Explainable Artificial Intelligence, Adversarial Attacks, Optimization**

I. Introduction

Deep learning has achieved remarkable success in critical time series forecasting (TSF) domains, including finance and healthcare [1]. However, the "black-box" nature of these models raises serious concerns regarding trust and accountability in high-stakes scenarios [2]. To bridge this gap, Explainable Artificial Intelligence (XAI) has emerged as a crucial component in TSF pipelines [3]. While attribution methods facilitate model debugging and user trust [4], recent studies suggest these explanations are vulnerable to malicious manipulations, posing secondary safety risks [5]. *Adversarial XAI* aims to generate adversarial examples that alter explanations while keeping the original prediction virtually unchanged [6].

While explanation vulnerability is well-scrutinized in computer vision, the TSF domain remains severely underexplored [7]. Existing attacks face two primary limitations in TSF. First, generic distance metrics disregard the unique spectral characteristics of time series; naive perturbations disrupt intrinsic trends and seasonality, rendering attacks detectable. Second, prior works predominantly employ white-box strategies tailored to specific explainers [8], [9]. This limits their applicability against diverse, proprietary forecasting architectures [10], leaving the field without a unified mechanism to understand why black-box models succumb to manipulation without explainer-specific knowledge.

To address this fundamental gap, we investigate the geometric roots of explanation instability. We propose a core geometric hypothesis: *The fragility of any attribution-based explanation relying on local linearity is fundamentally governed by the curvature of the model's decision manifold.* Because explainers (e.g., SHAP, LIME) assume local linearity, navigating the prediction iso-surface to regions of extreme curvature induces a collapse of this approximation. This invalidates the explanation without altering prediction results, offering a universal perspective transcending specific algorithms.

To validate this hypothesis in black-box TSF settings, we employ a spectro-temporal constrained optimization algorithm. We utilize finite differences to estimate decision boundary curvature, Natural Evolution Strategies (NES) to search for optimal perturbations, and Fast Fourier Transform (FFT) to strictly preserve spectral characteristics. Extensive experiments on real-world datasets across diverse architectures empirically confirm that maximizing this geometric objective significantly destabilizes explanations, verifying that high curvature drives explanation fragility.

The main contributions of this paper are as follows:

- We deconstruct the vulnerability of XAI from a geometric perspective, establishing a unified link between the decision manifold curvature and explainer robustness. We demonstrate that maximizing local curvature on the iso-surface is a sufficient condition to destabilize any explanation method based on linear assumptions. This offers a novel, explainer-independent perspective for evaluating XAI safety.
- We present the systematic study to reveal the geometric fragility of black-box explainers in the time series forecasting domain. Utilizing our gradient-free curvature probing algorithmframework (based on NES and spectral constraints), we confirm that mainstream TSF models possess regions of extremely high curvature on their decision boundaries. This discovery indicates that attackers can manipulate explanations without accessing explainer

gradients, exposing a significant latent risk in current time series XAI systems.

II. Related Work

A. Explainability in Time Series Forecasting

Post-hoc XAI methods, which aim to elucidate the behavior of already-trained, complex models without modifying their internal structure, have gained widespread adoption for their flexibility, generality, and applicability to state-of-the-art black-box predictors [11]. Most prominent post-hoc approaches belong to attribution-based methods, which assign relevance scores or importance values to time steps indicating their contribution to a specific model output [12]. Attribution-based methods themselves can be further divided into two primary categories based on their underlying principles: gradient-based methods and perturbation-based methods [13]. Gradient-based methods (e.g., Saliency Map [14], Integrated Gradients [15], SmoothGrad [16]) assess feature importance by analyzing model gradients, while perturbation-based methods (e.g., LIME [17], SHAP [18], LRP [19]) determine feature importance by altering features to observe model output changes.

B. Adversarial Attacks on Model Performance

Early research in adversarial machine learning predominantly focused on compromising Forecasting Performance. Seminal works by Szegedy et al. and Goodfellow et al. introduced methods like FGSM and PGD to induce classification errors via imperceptible perturbations [20]. In the context of time series, these strategies have been adapted to degrade the accuracy of regression and classification models by injecting noise that maximizes prediction error [21]. While effective in identifying model robustness issues, these attacks solely target the final output values, leaving the interpretability mechanisms unexamined.

C. Adversarial Attacks on Explanations

Recent research has pivoted towards evaluating the robustness of explanations themselves, leading to the development of Explainer-Specific Attacks. Ghorbani et al. [9] demonstrated that attribution maps could be manipulated by optimizing perturbations against the gradient of the explainer. Similarly, Slack et al. [8] devised scaffolding attacks specifically for perturbation-based methods like LIME and SHAP, utilizing biased classifiers that behave normally on the data manifold but adversarially off-manifold. They are typically tightly coupled with specific explanation algorithms, requiring white-box access or method-specific bias injection, which limits their generality [10]. Furthermore, blindly transferring these image-centric methodologies to temporal data often introduces high-frequency artifacts that violate the signal's spectral properties, rendering the adversarial perturbations conspicuous to standard anomaly detection filters.

D. Geometric Analysis of Explanations Vulnerability

A more fundamental perspective links adversarial vulnerability to the geometry of the decision manifold. Dombrowski et al. [22] mathematically derived that the fragility of saliency maps is a direct consequence of the high curvature (Hessian norm) of the prediction manifold. Although this provides a rigorous explanation for gradient instability, their analysis is theoretically strictly limited to gradient-based explanations in a white-box setting. For the broader spectrum of attribution methods (particularly perturbation-based approaches), the geometric determinants of their robustness remain obscure. Current research lacks a unified theoretical framework to determine whether a common geometric characteristic governs the vulnerability of these diverse explanation paradigms, leaving the root cause of their instability in black-box scenarios largely unidentified.

III. Methodology

A. Problem Formulation

We consider a multivariate time series forecasting task. Let $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N}$ denote a time series dataset, where each sample $\mathbf{x}_i \in \mathbb{R}^{T \times V}$ represents a historical observation series of length T with V variables. A pre-trained black-box forecasting model $f : \mathbb{R}^{T \times V} \to \mathbb{R}^{\tau \times V'}$ maps the historical input to a future prediction. For a given prediction, a post-hoc explanation method $g(\mathbf{x}, f)$ generates an attribution map $\mathbf{A} \in \mathbb{R}^{T \times V}$, where each element signifies the relevance of a specific time step and variable to the model's output.

Adversarial attacks on explanations are formally defined as a constrained optimization problem. In this context, an attack is rigorously formalized as finding a perturbation that satisfies a specific adversarial goal. The prevailing paradigm formulates this goal by explicitly targeting the outcome of the explanation. It defines the optimization objective as maximizing the divergence between the original and perturbed explanations while maintaining model output consistency. This standard optimization problem is expressed as [5], [6], [22]:

$$\begin{aligned} \max_{\boldsymbol{\delta}} \quad & \mathcal{D}\left(g(\mathbf{x}, f), g(\mathbf{x}+\boldsymbol{\delta}, f)\right) \\ \text{s.t.} \quad & \|f(\mathbf{x}) - f(\mathbf{x}+\boldsymbol{\delta})\|_\infty \leq \epsilon_{pred}, \\ & \|\boldsymbol{\delta}\|_\infty \leq \epsilon_{dist}, \end{aligned} \tag{1}$$

where $\mathcal{D}$ is a discrepancy metric (e.g., Top-k intersection rate) and ϵ_{pred} defines the allowable margin for prediction change.

Limitations. This formulation is fundamentally constrained by its **coupling with the explainer** g. Optimizing Eq. 1 requires querying or differentiating through specific explanation methods, rendering it computationally prohibitive for non-differentiable explainers or proprietary black-box settings. Furthermore, simple distance constraints treat time points independently, ignoring temporal correlations and often introducing unnatural artifacts that disrupt the intrinsic trends and seasonality of the data.

To address the coupling issue, we propose a paradigm shift: **attacking the model's geometry rather than the**

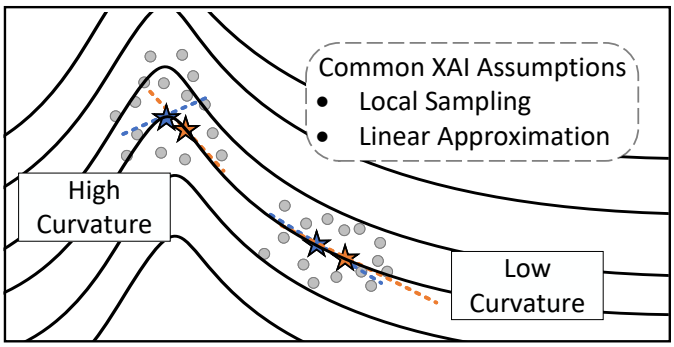


Fig. 1: Geometric intuition of explanation fragility. The black contours represent the decision iso-surfaces of the forecasting model. (Bottom Right) In low-curvature regions, the local linear surrogates (dashed lines) for the original point (blue star) and the perturbed point (orange star) are nearly parallel, indicating stable explanations. (Top Left) In high-curvature regions, infinitesimal changes in input position cause the local linear surrogates to intersect and rotate drastically. This geometric slope variance directly translates to the collapse of explanation robustness derived from perturbation-based methods.

explainer's output. We posit that the fragility of attribution-based explanations (regardless of whether they are gradient-based or perturbation-based) is intrinsically governed by the local curvature of the decision manifold. Accordingly, we replace the explainer discrepancy $\mathcal{D}$ with a **Geometric Proxy** $\mathcal{K}$. Additionally, to ensure the perturbation remains contextually valid in time series data, we impose a spectral constraint Ω_{freq}. The proposed optimization problem is:

$$\begin{aligned} \max_{\boldsymbol{\delta}} \quad & \mathcal{K}(f, \mathbf{x}+\boldsymbol{\delta}) \\ \text{s.t.} \quad & \|f(\mathbf{x}) - f(\mathbf{x}+\boldsymbol{\delta})\|_2 \leq \epsilon_{pred}, \\ & \boldsymbol{\delta} \in \Omega_{freq}. \end{aligned} \tag{2}$$

The critical innovation lies in substituting the explainer-dependent discrepancy $\mathcal{D}$ with the model-intrinsic curvature $\mathcal{K}$. This reformulation shifts the attack paradigm from manipulating the explanation output to exploiting the decision boundary's geometry. By directly maximizing curvature, we systematically invalidate the local linearity assumption essential to most explainers, allowing us to destabilize g without ever accessing or querying it. This achieves a truly universal and decoupled attack.

B. Theoretical Insight: Geometry of Explanation Fragility

The proposal to substitute the explainer-dependent discrepancy with a geometric curvature objective is grounded in a fundamental geometric property of attribution methods. In this section, we provide the theoretical justification for this substitution. We demonstrate that maximizing the curvature of the decision manifold is a sufficient condition to induce instability in attribution-based methods that relies on the assumption of local linearity.

Despite their algorithmic differences, post-hoc attribution methods share a unified mathematical premise: they assume that the complex, non-linear forecasting model f can be effectively approximated by a linear surrogate within a local neighborhood $\mathcal{N}(\mathbf{x})$ of the input $\mathbf{x}$. Our theoretical analysis targets this assumption.

For methods deriving attributions from gradients (e.g., Saliency Map, Integrated Gradients), the explanation is essentially $\nabla_{\mathbf{x}} f(\mathbf{x})$. It is well-established that the local stability of the gradient is governed by the Hessian matrix $\mathbf{H} = \nabla^2_{\mathbf{x}} f(\mathbf{x})$. Specifically, for a small perturbation $\boldsymbol{\delta}$, the shift in the explanation is given by the first-order Taylor expansion of the gradient:

$$\nabla f(\mathbf{x}+\boldsymbol{\delta}) - \nabla f(\mathbf{x}) \approx \mathbf{H}\boldsymbol{\delta}. \tag{3}$$

Thus, a decision manifold with high curvature (large spectral norm $\|\mathbf{H}\|_2$) inherently results in hypersensitive gradients, validating the curvature attack for this class of explainers [22].

A critical gap in existing literature is the lack of geometric analysis for perturbation-based methods (e.g., LIME, SHAP), which do not calculate gradients directly. We bridge this gap by proving that high curvature is equally destructive to the sampling-based linear surrogates used by these methods.

Perturbation-based methods generate explanations by fitting a linear model $g(\mathbf{z}) = \boldsymbol{\phi}^\top \mathbf{z}$ to approximate the model behavior $f(\mathbf{z})$ locally. This is typically formulated as a Weighted Least Squares (WLS) optimization:

$$\boldsymbol{\phi}^* = \arg\min_{\boldsymbol{\phi}} \sum_{\mathbf{z}\in\mathcal{N}(\mathbf{x})} \pi_{\mathbf{x}}(\mathbf{z}) \left(f(\mathbf{z}) - \boldsymbol{\phi}^\top \mathbf{z}\right)^2, \tag{4}$$

where $\pi_{\mathbf{x}}(\mathbf{z})$ is a proximity kernel (e.g., exponential kernel) defining the locality. The optimal solution $\boldsymbol{\phi}^*$ constitutes the explanation.

To analyze the impact of curvature, consider the second-order Taylor expansion of the k-th output component of the forecasting model $f_k(\mathbf{z})$ around $\mathbf{x}$:

$$f_k(\mathbf{z}) \approx f_k(\mathbf{x}) + \nabla f_k(\mathbf{x})^\top (\mathbf{z}-\mathbf{x}) + \frac{1}{2}(\mathbf{z}-\mathbf{x})^\top \mathbf{H}_k (\mathbf{z}-\mathbf{x}), \tag{5}$$

where $\mathbf{H}_k$ is the Hessian matrix corresponding to the k-th output dimension. The linear surrogate $\boldsymbol{\phi}^\top \mathbf{z}$ utilized by explanation methods fails to capture this quadratic component $\frac{1}{2}(\mathbf{z}-\mathbf{x})^\top \mathbf{H}_k (\mathbf{z}-\mathbf{x})$. Consequently, a large spectral norm $\|\mathbf{H}_k\|_2$ leads to a dramatic increase in the residual sum of squares (RSS) of the regression, destabilizing the explanation.

According to statistical regression theory, the variance of the estimated coefficients $\boldsymbol{\phi}^*$ is proportional to the variance of the residuals σ^2_{res}:

$$\text{Var}(\boldsymbol{\phi}^*) \propto \sigma^2_{res} \approx \mathbb{E}_{\mathbf{z}}[r(\mathbf{z})^2]. \tag{6}$$

Substituting the curvature dependence, we obtain:

$$\text{Var}(\boldsymbol{\phi}^*) \propto \|\mathbf{H}_k\|_F^2. \tag{7}$$

Based on the analysis above, we formally establish that the stability of perturbation-based explanations is inversely proportional to the curvature of the decision boundary. By

Algorithm 1 ICMA-Guided Adversarial Search

Input: Model f, Input $\mathbf{x}$, Steps K, Learning Rate α, Perturbation budget ϵ_{dist}, Prediction tolerance ϵ_{pred}.
Output: $\mathbf{x}_{adv}$. **Init:** $\boldsymbol{\delta} \leftarrow \mathbf{0}$.
for $t = 1$ to K **do**
  Sample $\{\mathbf{u}_i\}_{i=1}^{P} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and compute curvature costs J_i.
  $\mathbf{g} \leftarrow \frac{1}{P\sigma} \sum_{i=1}^{P} J_i(\mathbf{x} + \boldsymbol{\delta} + \sigma \mathbf{u}_i) \cdot \mathbf{u}_i$. // NES Estimation
  $\boldsymbol{\delta} \leftarrow \boldsymbol{\delta} + \alpha \cdot \text{sign}(\mathbf{g})$.
  $\boldsymbol{\delta} \leftarrow \text{Clip}(\mathcal{P}(\boldsymbol{\delta}), -\epsilon_{dist}, \epsilon_{dist})$. // Domain Constraints
  if $\|f(\mathbf{x} + \boldsymbol{\delta}) - f(\mathbf{x})\|_2 > \epsilon_{pred}$ **then**
    $\boldsymbol{\delta} \leftarrow \text{Backtrack}(\boldsymbol{\delta}, f, \epsilon_{pred})$. // Manifold Correction
  end if
end for
return $\mathbf{x}_{adv} \leftarrow \mathbf{x} + \boldsymbol{\delta}$.

navigating to a region where the Hessian norm $\|\mathbf{H}\|$ is maximized, we force the residual variance σ_{res}^2 to explode. This renders the linear approximation statistically insignificant, causing the resulting explanation $\boldsymbol{\phi}^*$ to fluctuate wildly or collapse.

This theoretical insight justifies our formulation in Eq. 2: explicitly maximizing curvature is a universal mechanism to attack both gradient and perturbation explainers without requiring access to the explainers themselves.

C. Optimization Strategy: ICMA-Guided Adversarial Search

We position our framework within a realistic strict black-box threat model. We assume the attacker has no knowledge of the forecasting model's internal parameters, gradients, or architecture, nor do they have access to the specific explanation algorithm being used. The attacker is limited solely to query access: submitting time series inputs and observing prediction outputs. This constraint precludes standard gradient-based attacks (e.g., PGD) and requires a methodology that operates purely on zeroth-order information.

To solve the curvature maximization problem under strict black-box conditions while preserving temporal semantics, we employ a spectro-temporal constrained derivative-free optimization strategy. First, we strictly confine perturbations to the high-frequency subspace using a differentiable projection operator $\mathcal{P}(\boldsymbol{\delta}) = \text{Re}(\mathcal{F}^{-1}(\mathbf{M} \odot \mathcal{F}(\boldsymbol{\delta})))$, where the high-pass mask $\mathbf{M}$ suppresses low-frequency trend modifications. Second, we approximate the intractable Hessian norm using a directional curvature estimator based on finite differences:

$$\mathcal{K}(\mathbf{z}) \approx \mathbb{E}_{\mathbf{u}} \left[\frac{\|f(\mathbf{z} + \sigma\mathbf{u}) - 2f(\mathbf{z}) + f(\mathbf{z} - \sigma\mathbf{u})\|_2}{\sigma^2} \right], \quad (8)$$

where $\mathbf{u}$ are random unit vectors. Finally, this objective is optimized via Natural Evolution Strategies (NES), which estimates gradients through population sampling to iteratively guide the input towards decision boundary regions with extreme geometric volatility using solely forward pass queries (detailed in Alg. 1).

D. Complexity Analysis

We analyze the computational efficiency of the proposed approach compared to standard explainer-coupled attacks. Let C_f denote the computational cost of a single inference pass of the forecasting model f.

Standard attacks that explicitly target the explanation discrepancy $\mathcal{D}$ must re-evaluate the explanation method g at every optimization step. For perturbation-based explainers like SHAP or LIME, generating a single explanation typically involves aggregating thousands of perturbed model queries to approximate local feature importance. If we denote this sampling size as M, the per-iteration complexity is $\mathcal{O}(M \cdot C_f)$, which is computationally expensive for iterative attacks.

In contrast, the proposed curvature-guided search targets the model geometry directly. The finite-difference curvature estimation via NES only requires a fixed population of model queries P per iteration. Consequently, the complexity is reduced to $\mathcal{O}(P \cdot C_f)$. Since the population size P is typically two orders of magnitude smaller than the explainer sampling size M, our approach is significantly more efficient, making it feasible for black-box robustness evaluation where direct explanation optimization is prohibitive.

IV. Experiments and Results

A. Experimental Setup

Datasets and Models. We use two real-world multivariate time series datasets: Beijing PM2.5 [1] (environmental, seasonal) and Shenzhen Stock [2] (financial, high volatility). We employ three forecasters: LSTM with input cell attention [23], TCN [24], and Transformer [25], trained for one-step-ahead prediction.

Target XAI Methods. We perform attacks against five popular attribution XAI methods, encompassing both gradient-based (Saliency Maps (SM) [14], Integrated Gradients (IG) [15], SmoothGrad (SG) [16]) and perturbation-based (LIME [17], and Kernel SHAP [18]) categories.

Baselines via Attack Objectives. To validate our geometric hypothesis, we compare three distinct optimization objectives: (1) **Random Noise**: Tests robustness against unstructured data deviations. (2) **Explanation Discrepancy (IFIA [22])**: Directly maximizes explanation discrepancy using gradient descent. This acts as a reference for differentiable explainers. (3) **Curvature Maximization (ICMA)**: Maximizes manifold curvature without accessing the explainer. Comparing ICMA against IFIA verifies whether geometric curvature serves as a valid proxy for explanation vulnerability.

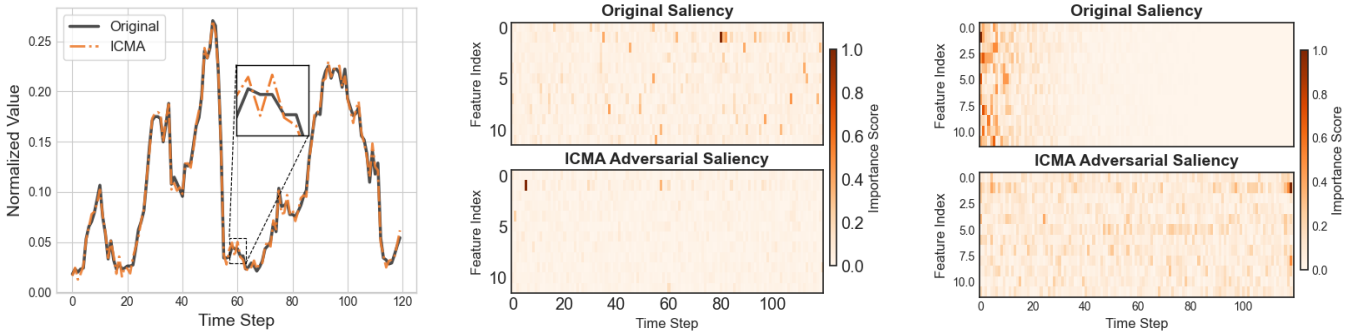
Evaluation Metrics & Implementation. We use Top-10% Intersection (TKI) [26] and Spearman's Rank Correlation (SRC) [27] to measure explanation robustness. Parameters are set to population size $P = 50$, iterations $K = 100$, and learning rate $\alpha = 0.005$. The frequency mask $\mathbf{M}$ zeroes out

[1] https://archive.ics.uci.edu/dataset/381/beijing+pm2+5+data

[2] The dataset was collected from the official website of the Shenzhen Stock Exchange by the authors, and the raw data was cleaned and stored in the GitHub repository. Our code is also in this repository. To respect the double-blind review policy, we exclude the url in manuscript.

TABLE I: Quantitative comparison of attack effectiveness on PM2.5 and Stock datasets.

		PM2.5						Stock					
		SRC			TKI			SRC			TKI		
		Random	IFIA	ICMA	Random	IFIA	ICMA	Random	IFIA	ICMA	Random	IFIA	ICMA
LSTM	LIME	0.99 ± 0.01	-	**0.68 ± 0.15**	0.95 ± 0.03	-	**0.74 ± 0.10**	1.00 ± 0.00	-	**0.67 ± 0.02**	0.98 ± 0.01	-	**0.68 ± 0.02**
	SHAP	0.99 ± 0.00	-	**0.62 ± 0.31**	0.97 ± 0.03	-	**0.68 ± 0.17**	1.00 ± 0.00	-	1.00 ± 0.00	0.98 ± 0.02	-	0.98 ± 0.02
	SM	1.00 ± 0.00	**0.81**	**0.81 ± 0.02**	0.88 ± 0.01	**0.90**	**0.90 ± 0.02**	1.00 ± 0.00	**0.80**	**0.80 ± 0.02**	1.00 ± 0.00	**0.90**	**0.90 ± 0.02**
	IG	0.61 ± 0.01	**0.54**	0.57 ± 0.01	0.80 ± 0.03	**0.73**	0.76 ± 0.03	0.99 ± 0.00	**0.98**	1.00 ± 0.00	0.97 ± 0.01	**0.93**	1.00 ± 0.00
	SG	1.00 ± 0.00	1.00	1.00 ± 0.00	0.97 ± 0.01	**0.96**	0.99 ± 0.00	1.00 ± 0.00	1.00	1.00 ± 0.00	1.00 ± 0.00	**0.99**	**0.99 ± 0.01**
TCN	LIME	0.93 ± 0.10	-	**0.14 ± 0.10**	0.88 ± 0.11	-	**0.27 ± 0.08**	0.90 ± 0.02	-	**0.65 ± 0.06**	0.80 ± 0.05	-	**0.65 ± 0.06**
	SHAP	0.95 ± 0.06	-	**0.31 ± 0.14**	0.90 ± 0.09	-	**0.45 ± 0.11**	0.98 ± 0.01	-	**0.94 ± 0.01**	0.93 ± 0.03	-	**0.87 ± 0.03**
	SM	0.76 ± 0.08	0.90	**0.86 ± 0.18**	0.79 ± 0.06	0.91	**0.88 ± 0.13**	0.85 ± 0.15	**0.08**	0.11 ± 0.23	0.81 ± 0.10	**0.36**	**0.36 ± 0.13**
	IG	0.90 ± 0.06	**0.26**	0.54 ± 0.08	0.78 ± 0.12	**0.42**	0.63 ± 0.06	0.89 ± 0.07	**-0.01**	0.07 ± 0.09	0.86 ± 0.06	0.35	**0.34 ± 0.04**
	SG	0.95 ± 0.00	0.57	**0.40 ± 0.05**	0.92 ± 0.01	0.65	**0.55 ± 0.04**	0.87 ± 0.04	**0.59**	0.60 ± 0.12	0.82 ± 0.04	**0.61**	0.64 ± 0.08
Transformer	LIME	0.99 ± 0.01	-	**0.52 ± 0.10**	0.96 ± 0.03	-	**0.60 ± 0.09**	0.96 ± 0.04	-	**0.60 ± 0.06**	0.92 ± 0.05	-	**0.62 ± 0.06**
	SHAP	0.99 ± 0.01	-	**0.68 ± 0.06**	0.97 ± 0.03	-	**0.68 ± 0.07**	1.00 ± 0.00	-	**0.96 ± 0.01**	0.99 ± 0.02	-	**0.91 ± 0.03**
	SM	0.98 ± 0.12	0.97	**0.92 ± 0.16**	0.96 ± 0.10	0.95	**0.91 ± 0.14**	0.84 ± 0.15	**0.33**	0.34 ± 0.09	0.85 ± 0.15	**0.52**	**0.52 ± 0.09**
	IG	0.94 ± 0.02	**0.08**	0.36 ± 0.12	0.90 ± 0.02	**0.34**	0.46 ± 0.09	0.93 ± 0.04	**0.25**	0.44 ± 0.03	0.95 ± 0.03	**0.47**	0.56 ± 0.03
	SG	0.97 ± 0.00	0.78	**0.77 ± 0.10**	0.90 ± 0.02	**0.75**	0.76 ± 0.08	0.71 ± 0.04	0.33	**0.28 ± 0.07**	0.71 ± 0.03	**0.49**	0.61 ± 0.03



(a) Samples (b) Explanations-SHAP (c) Explanations-LIME

Fig. 2: Visualization of ICMA on the PM2.5 dataset with a Transformer model.

the lowest 10% of spectral coefficients to preserve trends. And we set $\epsilon_{pred} = 0.01$ and $\epsilon_{dist} = 0.05$ in our experiments.

B. Analysis of Attack Effectiveness

Table I reports the mean quantitative results over 10 runs on the PM2.5 and Stock datasets.

Baseline Comparison. The Random Noise baseline consistently yields high metrics, indicating that the original explanations are robust to unstructured perturbations. This comparison confirms that the significant reduction in explanation robustness achieved by our method results from the directed optimization of geometric curvature rather than random data artifacts.

Curvature as an Optimization Proxy. We compare our approach against IFIA, which directly maximizes explanation discrepancy using white-box gradients. On complex architectures such as TCN and Transformer, our curvature-driven method achieves attack success rates comparable to IFIA. This similarity indicates that maximizing the local curvature of the decision manifold is functionally equivalent to directly optimizing the explanation error. Consequently, curvature serves as an effective proxy for evaluating explanation robustness without requiring access to the explanation model itself.

Vulnerability of Perturbation-based Explainers. Our method demonstrates effectiveness against Perturbation-based methods (LIME, SHAP) where IFIA is inapplicable. These results confirm that perturbation-based methods, which rely on local linear approximations, fail in the high-curvature regions identified by our search, even when the attacker does not have access to the specific explainer.

Architecture and XAI Sensitivity. The results reveal a distinct hierarchy of vulnerability driven by model topology: TCN $>$ Transformer $>$ LSTM. TCNs exhibit the highest sensitivity, as their dilated convolutional structures tend to create sharp, highly non-linear decision boundaries with high local curvature, which our attack easily exploits. In contrast, LSTMs demonstrate superior robustness; their gating mechanisms naturally saturate, flattening the decision manifold and dampening the local curvature required to destabilize explanations.

Regarding XAI methods, perturbation-based methods prove highly susceptible, confirming that high-curvature regions successfully invalidate their core assumption of local linearity. Among gradient-based methods, SG shows notable resilience. This is geometrically consistent, as SG averages gradients over Gaussian noise, effectively acting as a low-pass filter

that smooths out the local curvature spikes introduced by the attack.

C. Visualization Analysis

To intuitively validate the method, Fig. 2 visualizes the results on the PM2.5 dataset using the Transformer model. The figure contrasts the time-domain samples with the corresponding interpretations for SHAP and LIME.

In Fig. 2(a), the adversarial sample (orange dashed line) closely follows the trend of the original sequence (black solid line). The perturbations are restricted to the high-frequency subspace, preserving the underlying temporal semantics. In contrast, the explanation heatmaps in Fig. 2(b) and (c) show significant divergence. The original feature importance patterns are replaced by dispersed noise or shifted focus areas in the adversarial counterparts. This demonstrates that high local curvature leads to substantial variations in explanation results while maintaining prediction and visual consistency.

V. Conclusion

This work identifies the curvature of the decision manifold as the fundamental determinant of explanation robustness, demonstrating that maximizing curvature is sufficient to destabilize diverse black-box explainers. This geometric perspective unifies the understanding of XAI vulnerability, shifting the safety focus from specific explainer algorithms to the intrinsic topology of forecasting models. Future work will explore geometric defenses to mitigate these latent risks.

Acknowledgment

This work was patially supported by the Anhui Provincial Science and Technology Breakthrough Project under Grant No, 202523o09050015 and Advanced Micro-Fabrication Equipment Inc. China (AMEC) Innovative Research under Grant No. EF2190000055.

References

[1] B. Lim and S. Zohren, "Time-series forecasting with deep learning: a survey," *Philosophical transactions of the royal society a: mathematical, physical and engineering sciences*, vol. 379, no. 2194, 2021.

[2] T.-C. T. Chen, "Applications of xai for forecasting in the manufacturing domain," in *Explainable artificial intelligence (XAI) in manufacturing: methodology, tools, and applications*. Springer, 2023, pp. 13–50.

[3] C. Van Zyl, X. Ye, and R. Naidoo, "Harnessing explainable artificial intelligence for feature selection in time series energy forecasting: A comparative analysis of grad-cam and shap," *Applied Energy*, vol. 353, p. 122079, 2024.

[4] P. F. Pérez, C. Fiandrino, E. P. Gómez, H. Mohammadalizadeh, M. Fiore, and J. Widmer, "Aichronolens: Ai/ml explainability for time series forecasting in mobile networks," *IEEE Transactions on Mobile Computing*, vol. 24, no. 8, pp. 7757–7772, 2025.

[5] F. Leofante and M. Wicker, *Robust Explainable AI*. Springer, 2025.

[6] H. Baniecki and P. Biecek, "Adversarial attacks and defenses in explainable artificial intelligence: A survey," *Information Fusion*, vol. 107, p. 102303, 2024.

[7] A. Jyoti, K. B. Ganesh, M. Gayala, N. L. Tunuguntla, S. Kamath, and V. N. Balasubramanian, "On the robustness of explanations of deep neural network models: A survey," *arXiv preprint arXiv:2211.04780*, 2022.

[8] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju, "Fooling lime and shap: Adversarial attacks on post hoc explanation methods," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2020, pp. 180–186.

[9] A. Ghorbani, A. Abid, and J. Zou, "Interpretation of neural networks is fragile," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, pp. 3681–3688, 2019.

[10] W. Huang, X. Zhao, G. Jin, and X. Huang, "Safari: Versatile and efficient evaluations for robustness of interpretability," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 1988–1998.

[11] C. O. Retzlaff, A. Angerschmid, A. Saranti, D. Schneeberger, R. Roettger, H. Mueller, and A. Holzinger, "Post-hoc vs ante-hoc explanations: xai design guidelines for data scientists," *Cognitive Systems Research*, vol. 86, p. 101243, 2024.

[12] T. Rojat, R. Puget, D. Filliat, J. Del Ser, R. Gelin, and N. Díaz-Rodríguez, "Explainable artificial intelligence (xai) on timeseries data: A survey," *arXiv preprint arXiv:2104.00950*, 2021.

[13] L. Qiu, Y. Yang, C. C. Cao, Y. Zheng, H. Ngai, J. Hsiao, and L. Chen, "Generating perturbation-based explanations with robustness to out-of-distribution data," in *Proceedings of the ACM Web Conference 2022*, ser. WWW '22. Association for Computing Machinery, 2022, p. 3594–3605.

[14] D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, and K.-R. Müller, "How to explain individual classification decisions," *The Journal of Machine Learning Research*, vol. 11, pp. 1803–1831, 2010.

[15] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *International conference on machine learning*. PMLR, 2017, pp. 3319–3328.

[16] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "Smoothgrad: removing noise by adding noise," *arXiv preprint arXiv:1706.03825*, 2017.

[17] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?": Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 1135–1144.

[18] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 4768–4777.

[19] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PloS one*, vol. 10, no. 7, p. e0130140, 2015.

[20] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[21] H. I. Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.-A. Muller, "Adversarial attacks on deep neural networks for time series classification," in *2019 International joint conference on neural networks (IJCNN)*. IEEE, 2019, pp. 1–8.

[22] A.-K. Dombrowski, M. Alber, C. Anders, M. Ackermann, K.-R. Müller, and P. Kessel, "Explanations can be manipulated and geometry is to blame," *Advances in neural information processing systems*, vol. 32, 2019.

[23] A. A. Ismail, M. Gunady, L. Pessoa, H. Corrada Bravo, and S. Feizi, "Input-cell attention reduces vanishing saliency of recurrent neural networks," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[24] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*, 2018.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.

[26] Z. Wang, H. Wang, S. Ramkumar, P. Mardziel, M. Fredrikson, and A. Datta, "Smoothed geometry for robust attribution," *Advances in neural information processing systems*, vol. 33, pp. 13 623–13 634, 2020.

[27] T. Fel, D. Vigouroux, R. Cadène, and T. Serre, "How good is your explanation? algorithmic stability measures to assess the quality of explanations for deep neural networks," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 720–730.